ScamX — 실시간 AI 스캠 탐지 앱 기술서

1. 문서 소개

이 문서는 ScamX 프로젝트의 기획 배경, 기술적 의사결정, Android 정책 대응, AI 파이프라인 설계, 개인정보 보호 전략, 그리고 실제 개발 과정에서 겪은 문제와 해결 과정을 정리한 기술 보고서다.
ScamX 는 단순한 스팸 필터가 아니라, **문자 내용을 의미 기반으로 분석해 신종 스캠까지 탐지하는 실시간 보안 앱**을 목표로 설계되었다.

2. 프로젝트 개요

ScamX 는 Android 환경에서 스캠 문자를 실시간으로 탐지하는 개인 프로젝트다.
기획, 백엔드, 프론트엔드, 모바일 개발, AI 파이프라인 설계까지 전부 직접 수행했다.

핵심 목표는 다음과 같다.

- 통신사 필터링으로 잡히지 않는 가족·지인 사칭형 스캠 탐지
- 패턴 기반이 아닌 의미 기반 스캠 분석
- Android 10+ 정책을 준수하는 합법적 문자 수집 구조
- 개인정보를 저장하지 않는 비식별화 AI 분석 구조
- 사용할수록 정확도가 올라가는 Self-Improving 시스템

3. 문제 인식

ScamX 를 만들게 된 배경은 다음과 같다.

- 통신사 스팸 필터는 발신번호 기반이라 저장된 번호로 위장한 스캠을 잡지 못함
- 기존 보안 앱은 패턴 DB 매칭 방식이라 신종 스캠에 취약함
- Android 10 이후 SMS 브로드캐스트 방식이 금지되어 문자 본문을 직접 읽을 수 없음
- 고령층은 스캠에 취약하지만 기존 앱들은 사용성이 복잡함

이 문제들을 해결하기 위해 **문자 내용을 의미적으로 분석하는 보안 앱**이 필요하다고 판단했다.

4. Android 정책 대응

처음에는 기존 보안 앱처럼 SMS 브로드캐스트로 본문을 읽으려 했지만, Android 10 이후 Google 정책이 강화되면서 이 방식이 완전히 금지된 것을 확인했다.
메시지 앱보다 먼저 SMS 본문을 읽는 방식은 정책 위반이며, 개인정보보호법과도 충돌한다.

이 문제를 해결하기 위해 ScamX 는 **Notification Listener 기반으로 문자 본문을 읽는 방식**을 채택했다.
이 방식은 정책 위반이 아니며, 모든 Android 기기에서 안정적으로 동작한다.

5. 개인정보 보호 설계

문자 본문을 서버로 보내는 구조이기 때문에, 초기 설계부터 **원문을 절대 저장하지 않는 구조**를 목표로 했다.

ScamX 는 다음과 같은 3 단계 마스킹 파이프라인을 사용한다.

- 서버 도착 즉시 원문 마스킹
- 전화번호는 해시 처리
- 계좌번호·이름 등은 패턴 기반 마스킹
- 민감 단어는 제거 또는 대체
- AI 분석 후에는 판단 결과만 사용자에게 전달
- DB 에는 원문이 아닌 비식별화된 의미 벡터만 저장

즉, 서버에는 문자 내용이 아니라 **문자 의미만 남는다.**

6. 아키텍처 개요

ScamX 는 다음과 같은 구조로 구성되어 있다.

- 모바일: Kotlin 기반 Android 앱, Notification Listener 로 문자 수집
- 백엔드: Python, FastAPI, LangGraph 기반 AI 파이프라인
- AI 모델: KOLLECTRA, Sentence Embedding, LLaMA 3.1 8B, GPT-5 Mini
- 데이터베이스: PostgreSQL(NeonDB), pgvector
- 프론트엔드: Next.js 기반 공공기관용 어드민 페이지

7. AI 파이프라인 설계

7.1 KOLLECTRA – 1차 스캠 필터링

SMS 알림 수신 직후 가장 먼저 동작하는 단계다.
경량 Transformer 모델로 스캠/정상 확률을 빠르게 계산해 정상 메시지는 여기서 종료한다.
이 단계 덕분에 불필요한 LLM 호출을 대폭 줄일 수 있었다.

7.2 Sentence Embedding – 의미 벡터 생성

KOLLECTRA가 "의심"으로 판단한 메시지는 임베딩 단계로 넘어간다.
문장을 768~1024차원 의미 벡터로 변환해 pgvector에 저장된 기존 스캠 벡터와 비교한다.
스캠은 문구만 조금씩 바꿔 반복되기 때문에 의미 기반 비교가 필수적이다.

7.3 LLaMA 3.1 8B – 의미 기반 스캠 분석

유사도만으로 판단이 어려운 문장은 LLaMA가 의미적 맥락을 기반으로 분석한다.
데이터가 쌓일수록 판단 정확도가 올라가는 구조다.

7.4 GPT-5 Mini – 신종 스캠 정밀 분석

LLaMA가 판단을 내리지 못하는 경우에만 GPT-5 Mini가 동작한다.
완전히 새로운 유형의 스캠을 분석하고, 결과는 마스킹 후 DB에 저장된다.
이후부터는 LLaMA가 처리할 수 있어 Self-Improving 구조가 완성된다.

8. 전체 파이프라인 흐름

SMS 알림 수신
→ KOLLECTRA 1차 필터링
→ Sentence Embedding 생성
→ pgvector 유사도 검색
→ LLaMA 의미 분석
→ 필요 시 GPT-5 Mini 정밀 분석
→ 결과 마스킹 후 DB 저장
→ 사용자 알림

9. UX 설계 – 고령층 고려

테스트 과정에서 고령층이 알림을 놓치는 문제가 반복적으로 발생했다.
이를 해결하기 위해 다음과 같은 UX 개선을 적용했다.

- 알림이 상단바에 일정 시간 유지
- 알림 센터에서 재확인 가능
- 클릭 시 상세 분석 페이지로 이동
- 알림 지속 시간 옵션 제공 예정

고령층이 놓치지 않도록 여러 단계에서 확인할 수 있는 구조로 설계했다.

10. 개발 과정에서 가장 어려웠던 점

ScamX 개발에서 가장 힘들었던 부분은 기술적인 문제보다 **정책과 현실적인 제약을 맞추는 과정**이었다.

- Android 10 정책으로 기존 방식이 전부 막혀 있었고
- 개인정보보호법 때문에 서버 구조를 완전히 새로 설계해야 했고
- LLM 비용이 너무 높아 파이프라인을 여러 번 갈아엎었고
- 스캠 문구는 계속 변해 단순 패턴 매칭이 통하지 않았다

이 과정에서 Notification Listener 기반 수집,
KOLLECTRA → 임베딩 → LLaMA → GPT 구조,
의미 벡터 기반 Self-Improving 시스템이 완성되었다.

11. 성과 및 차별점

- LLM 호출 최소화로 API 비용 약 70% 절감
- 신종 스캠 자동 축적 → 재학습 없이 정확도 향상
- Android 10+ 정책 완전 준수
- 공공기관 연동 가능한 어드민 페이지 구축
- 고령층도 알림 한 번으로 위험 여부 즉시 인지 가능

12. 한 줄 요약

의미 기반 분석으로 신종 스캠까지 잡아내는 실시간 AI 스캠 탐지 앱 – ScamX